Poisson regression: Further topics

Patrick Breheny

April 21

Overdispersion

- One of the defining characteristics of Poisson regression is its lack of a scale parameter: $\mathrm{E}(Y) = \mathrm{Var}(Y)$, and no parameter is available to adjust that relationship
- In practice, when working with Poisson regression, it is often the case that the variability of y_i about $\hat{\lambda}_i$ is larger than what $\hat{\lambda}_i$ predicts
- This implies that there is more variability around the model's fitted values than is consistent with the Poisson distribution

Overdispersion (cont'd)

- The term for this phenomenon is *overdispersion*
- Data for which this phenomenon manifests itself are often called "overdispersed", although as we will see, it is perhaps better to refer to the model as overdispersed, not the data
- There are two common approaches to correcting for overdispersion:
  - Quasi-likelihood
  - Negative binomial regression

Tinkering with the score

- Recall that the score arising from a Poisson regression model is

$$\frac{\partial \ell}{\partial \theta} = \sum_i \{y_i - \hat{\lambda}_i\}$$

  where $\theta = \log(\lambda)$, the canonical parameter
- Note, of course, that there is no scale parameter, which would show up in the denominator on the right hand side
- Now suppose we add one:

$$\frac{\partial \ell}{\partial \theta} = \sum_i \frac{y_i - \hat{\lambda}_i}{\phi}$$

Implications of our tinkering

- Recall that $\mathrm{Var}(Y) = \phi V(\mu)$; thus, we now have a parameter that allows the variance to be larger or smaller than the mean by a multiplicative factor ϕ
- This will not change $\widehat{\boldsymbol{\beta}}$, of course
- However, it will affect inference, since

$$\widehat{\boldsymbol{\beta}} \dot{\sim} \mathrm{N}\left(\boldsymbol{\beta}, \phi(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\right)$$

Quasi-likelihood

- So what distribution is this, that gives rise to this score?
- There isn't one (at least, not one for which you can write down the distribution in closed form)
- This approach, where you modify the score directly and never actually specify a distribution, is known as *quasi-likelihood*

Quasi-likelihood: Estimation of scale

- Typically, the scale parameter ϕ is estimated using the method of moments estimator

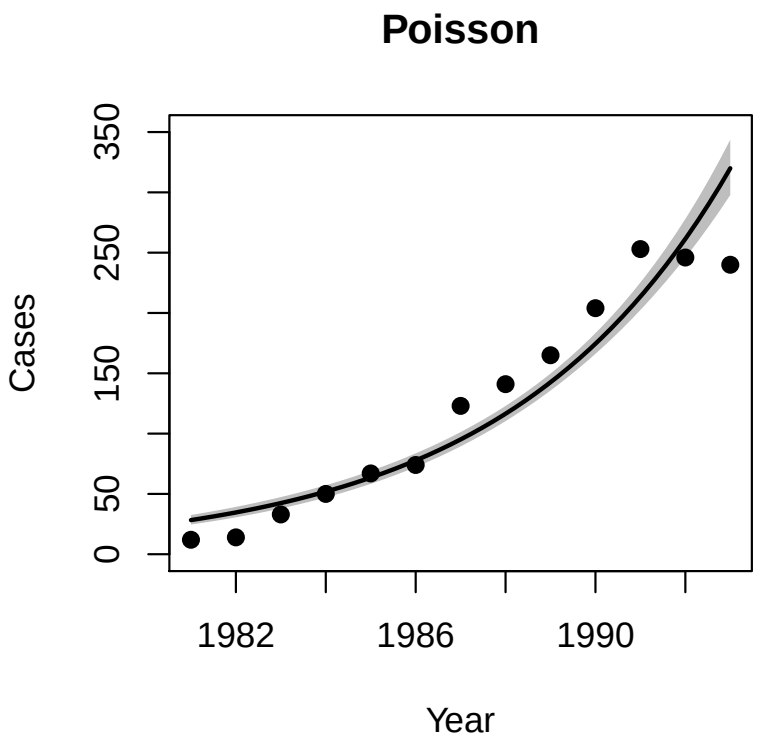
$$\hat{\phi} = \frac{X^2}{n-p}$$

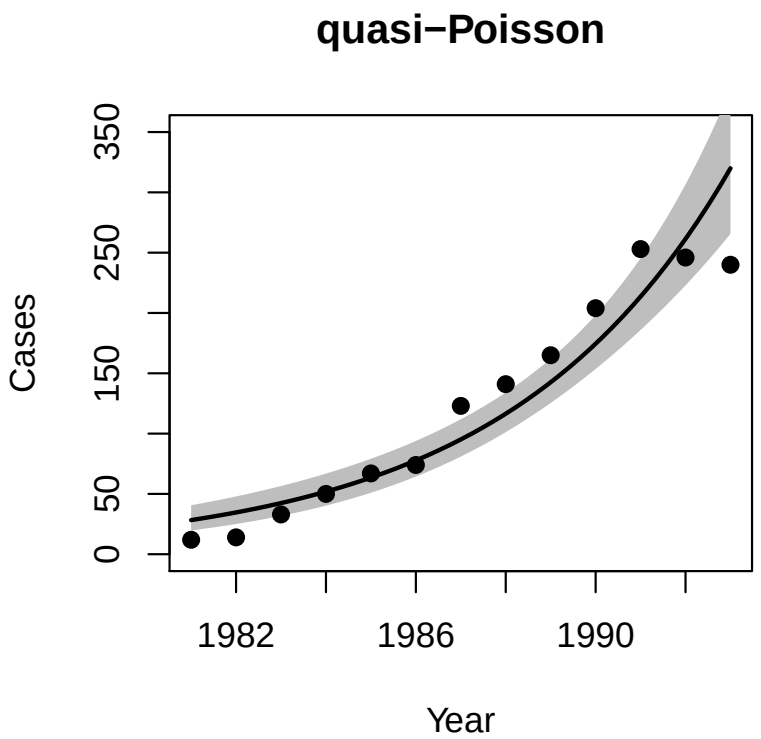
- To use this approach in `R`, one can specify `family=quasipoisson`; in SAS, one can add a `PSCALE` option to the model statement

Quasi-likelihood: Belgian AIDS data

- For our Belgian AIDS data, $\hat{\phi} = 6.7$, implying that the variance was nearly 7 times larger than that implied by the Poisson distribution
- Again, the fit is the same
- However, our standard errors are $\sqrt{6.7} \approx 2.6$ times larger

Quasi-likelihood: Belgian AIDS data (cont'd)





Drawbacks of quasi-likelihood

- The quasi-Poisson approach is attractive for several reasons, but its big drawback is that lacks a log-likelihood
- This prevents you from using any of the likelihood-based tools we have discussed for GLMs: likelihood ratio tests, AIC/BIC, deviance explained, deviance residuals
- An alternative approach that allows all those maximum likelihood tools is based on the negative binomial distribution

The negative binomial distribution

- The negative binomial distribution has other uses in probability and statistics, but for our purposes we can think about it as arising from a two-stage hierarchical process:

$$Z \sim \text{Gamma}(\theta, \theta)$$
$$Y|Z \sim \text{Poisson}(\lambda Z)$$

- The marginal distribution of Y is then negative binomial, with

$$\text{E}(Y) = \lambda$$
$$\text{Var}(Y) = \lambda + \lambda^2/\theta$$

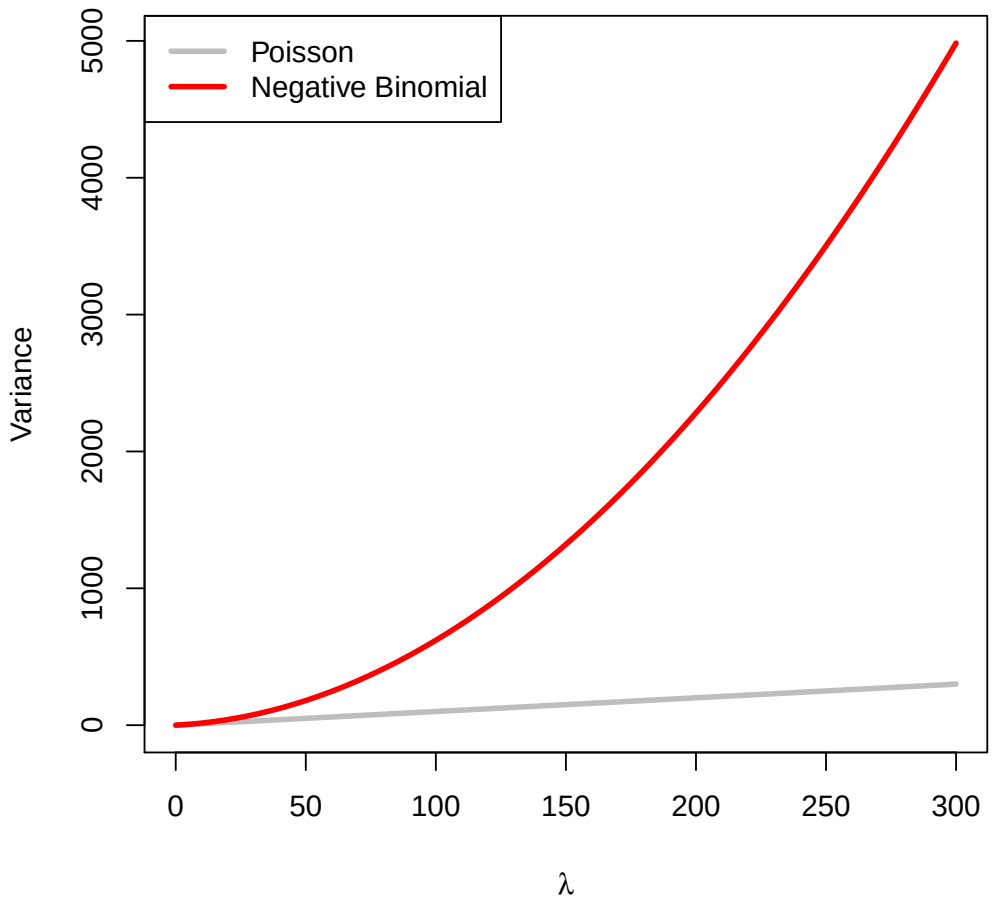
- Thus, like the Poisson distribution, the negative binomial has support only on the positive integers, but unlike the Poisson, its variance is larger than its mean

Negative binomial and exponential family

- Note, however, that the negative binomial distribution is not a member of the exponential family
- Thus, the theory and fitting procedures we have developed for GLMs do not directly apply here
- For example, there is no "canonical link"; however, it is customary to employ a log link to make negative binomial regression look like Poisson regression

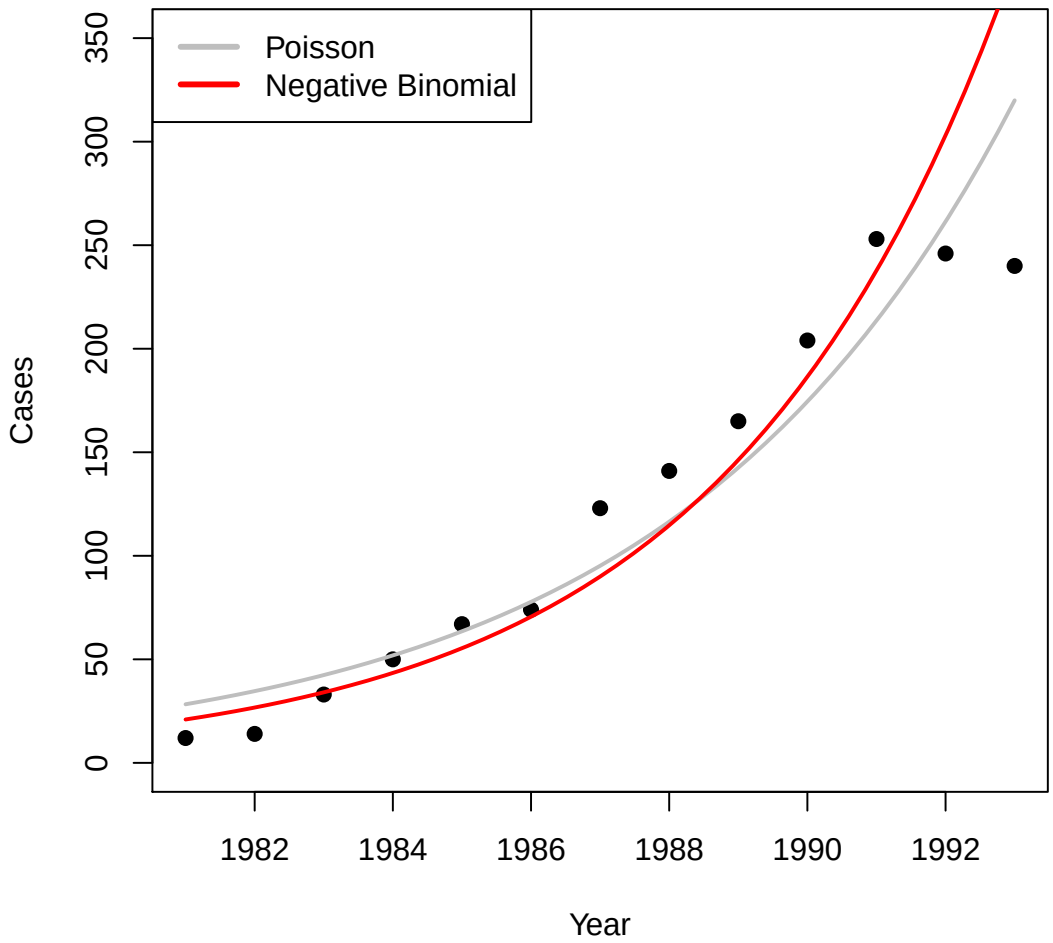
Negative binomial: Mean-variance relationship

For the Belgian AIDS data, $\hat{\theta} = 19.2$, implying the following mean-variance relationship:



Negative binomial: Estimate

This leads to the following estimate:



Remarks

- By any reasonable assessment, the negative binomial estimates here are worse than the Poisson fit – and certainly drastically worse than the quadratic Poisson model
- However, its "goodness of fit" measures are much better
- This is why I remarked earlier that it's wrong to think of the *data* as overdispersed – if the data show more variability than the model can explain, the most likely explanation is a bad model
- The quadratic Poisson fit shows no overdispersion (the residuals are actually slightly "underdispersed")

Remarks (cont'd)

- The key concept here is that residual variance is caused by two things: random variation and systematic bias in the model
- Many analysts have the mistaken view that quasi-Poisson or negative binomial regression "automatically" fixes the overdispersion problem
- This is a dangerous misconception – systematic bias in the model should take far greater priority than modeling the random error
- Quasi-Poisson or negative binomial should be thought of more as a last resort to fixing overdispersion – the first step is fixing the systematic component of the model

Poisson regression for contingency tables

- Another use for Poisson regression is to analyze contingency tables
- Recall the results of the Salk vaccine trial:

	Size of group	Polio cases per 100,000 children
Treatment	200,000	28
Control	200,000	71

Logistic vs. Poisson

One may consider two sorts of GLMs for this data:

- A logistic regression model, in which

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \texttt{Treatment}$$

- A Poisson regression model, in which

$$\log(\lambda_i) = \beta_0 + \beta_1 \texttt{Treatment}$$

Logistic vs. Poisson (cont'd)

Comparing our two estimates (the odds ratio for the logistic regression model and the rate ratio for the Poisson model), we see that they are exactly the same:

			95% CI		
Quantity	Model	Estimate	Lower	Upper	p
Rate ratio	Poisson	2.54	1.87	3.48	5.08×10^{-10}
Odds ratio	Logistic	2.54	1.87	3.48	5.08×10^{-10}

Multiple categories

- This is an interesting result to be aware of, as the Poisson distribution is more readily extended to multiple outcome categories than the binomial distribution is
- For example, our textbook contains the following data from a study of a new influenza vaccine, where the outcome was antibody levels, categorized as small/moderate/large:

	Antibody levels		
	Small	Moderate	Large
Placebo	25	8	5
Vaccine	6	18	11

Testing for association

- We can model these six counts as Poisson random variables with offset $n_0 = 38$ for the placebo group and $n_1 = 35$ for the vaccine group, then test the null hypothesis that the small/moderate/large rates are the same for the vaccine group as they are for the placebo group
- Assuming we parameterize the model in the usual way, this amounts to a test of the interaction term between antibody levels and group
- A likelihood ratio test of the full model in which each cell has its own Poisson rate vs. the restricted model in which the rates are the same in each group points is highly significant ($p = 0.00009$), indicating that we are unlikely to have seen such a large antibody response in the vaccine group due to chance alone

Remarks

- In this case, we could just have used a χ^2 or Fisher's Exact Test to accomplish the same thing
- The advantage of the Poisson model in general is that it allows us to build more complicated models with additional explanatory variables, and to model continuous variables using linear trends
- In practice, these models become unwieldy rather quickly as we try to add complexity
- Next time, we'll talk about ways to extend logistic regression to the multi-category case